Uncertainty quantification for Bayesian nonparametric estimators of rare species variety

Uncertainty quantification for Bayesian nonparametric estimators of rare species variety

Quantificazione dell'incertezza per stimatori bayesiani nonparametrici del numero di specie rare

Stefano Favaro and Bernardo Nipoti

Abstract Species sampling problems have regained popularity in recent years due to their frequent appearance in challenging applications arising from ecology, genetics, linguistic, etc. Interest often lies in estimating the number of rare species that appear in a sample, that is species with a frequency smaller than a specific abundance threshold. The Bayesian nonparametric approach has proved successful by providing closed form estimators for rare species variety. In this paper we present a novel methodology for endowing such estimators with asymptotic credible intervals. We illustrate it through the analysis of some genomic datasets.

Abstract *Il campionamento di specie è un tema di ricerca che, di recente, ha raccolto un rinnovato interesse per via della sua comparsa in applicazioni legate all'ecologia, la genetica, la linguistica, etc. Spesso l'interesse si concentra sulla stima del numero di specie rare, cioè specie con frequenza minore di una certa soglia, che compaiono in un campione. L'approccio Bayesiano nonparametrico si è dimostrato uno strumento potente che ha portato ad ottenere espressioni in forma chiusa per stimatori del numero di specie rare. In questo lavoro proponiamo una nuova metodologia per associare intervalli di credibilità a tali stimatori e la illustriamo con l'analisi di una dateset genomico.*

Key words: Species sampling problems, Poisson-Dirichlet process, Uncertainty quantification, Asymptotic credible intervals, Rare species.

Stefano Favaro
Department of Economics and Statistics, University of Torino, c.so Unione Sovietica 218/bis, 10134 Torino,e-mail: stefano.favaro@unito.it

Bernardo Nipoti
Department of Economics and Statistics, University of Torino, c.so Unione Sovietica 218/bis, 10134 Torino, e-mail: bernardo.nipoti@unito.it

1 Introduction

We consider the framework where an experimenter is sampling from a population of individuals belonging to different species with unknown proportions. Assuming an ideally infinite number of species, species labels are denoted by $(X_i^*)_{i\geq 1}$ and their respective proportions in the population by $(p_i)_{i\geq 1}$. Given an initial sample of size n, the object of our study is $M_{n,m}(l)$, the number of species with specific abundance l in an enlarged sample of size $(n+m)$, where the additional sample of size m has not been observed. More specifically, estimation of rare species variety consists in determining the number of species for which l is not greater than a certain abundance threshold τ.

The estimation of rare species variety has regained popularity in recent years due to its frequent appearance in challenging applications arising from a wide range of fields such as ecology, biology, bioinformatics, genetics, linguistic, etc. As a remarkable example, in genetics one is interested in estimating the number of individuals with rare genes, the reasons being that rare genes of a type may be associated with a deleterious disease. We refer to [2] for a detailed account on this topic.

1.1 Bayesian nonparametric estimators

Recent literature (see [3], [4]) has proved that the Bayesian nonparametrics approach provides a powerful tool which naturally leads to a simple and exact expression for the estimator $\hat{M}_{n,m}(l)$. Such approach, first introduced in [6], is based on the randomization of the unknown species proportions p_i's. Specifically, let $\tilde{P} = \sum_{i\geq 1} p_i \delta_{X_i^*}$ be a discrete random probability measure, namely $(p_i)_{i\geq 1}$ are non-negative random weights with some distribution such that $\sum_{i\geq 1} p_i = 1$ almost surely, and $(X_i^*)_{i\geq 1}$ are random locations independent of $(p_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic probability measure ν_0. A sample of n individuals $(X_1,\ldots,X_n)$ is taken from a population with composition directed by $\tilde{P}$, namely

$$\begin{aligned} X_i \,|\, \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \qquad i = 1,\ldots,n \\ \tilde{P} &\sim \Pi, \end{aligned} \tag{1}$$

for any $n \geq 1$, with Π playing the role of the prior. Then, according to the de Finetti representation theorem, $(X_i)_{i\geq 1}$ is an exchangeable sequence. Although in literature the number $M_{n,m}(l)$ has been studied under more general settings, for the sake of simplicity hereafter we assume that Π is the distribution of a two parameter Poisson-Dirichlet process that we denote by $\tilde{P}_{\sigma,\theta}$, with $\sigma \in (0,1)$ and $\theta > -\sigma$ (see [7] and [8] for details).

Let $K_n = j \leq n$ be the number of distinct species featured by the sample $(X_1,\ldots,X_n)$ with corresponding frequencies $(M_{n,0}(1),\ldots,M_{n,0}(n)) = (m_1,\ldots,m_n)$ such that

$\sum_{i=1}^{n} M_{n,0}(i) = K_n$ and $\sum_{i=1}^{n} iM_{n,0}(i) = n$. Moreover, let $(X_{n+1}, \ldots, X_{n+m})$ be an additional unobserved sample of size $m \geq 1$. A Bayesian nonparametric estimator of the number $K_m^{(n)}$ of new distinct species in the additional unobserved sample was derived in [6] and [3], whereas [4] determined the Bayesian nonparametric estimator for $M_{n,m}(l)$, with $l \in \{1, \ldots, n+m\}$. We can write, for any $l = 1, \ldots, n+m$,

$$\hat{M}_{n,m}(l) = \sum_{i=0}^{l} \binom{m}{l-i} m_i (i-\sigma)_{(l-i)} \frac{(\theta+n-i+\sigma)_{(m-l+i)}}{(\theta+n)_{(m)}}, \tag{2}$$

where, for simplifying the notation, we agree that $m_0 := -(\theta + \sigma j)/\sigma$ and we denote by $(a)_q = \Gamma(a+q)/\Gamma(a)$ the q-th ascending factorial of a.

Moreover one has that, for every $l \in \{1, \ldots, n+m\}$,

$$\frac{M_{n,m}(l)}{m^{\sigma}} \xrightarrow{\text{w}} \frac{\sigma(1-\sigma)_{l-1}}{l!} S_{\sigma,\theta,n,j} \tag{3}$$

where

$$S_{\sigma,\theta,n,j} \stackrel{\text{d}}{=} B_{j+\theta/\sigma, n/\sigma-j} S_{\sigma,(\theta+n)/\sigma}, \tag{4}$$

$B_{a,b}$ being a random variable distributed according to Beta distribution with parameter (a,b), and $S_{\sigma,\theta}$ a random variable with density function

$$f_{S_{\sigma,q}}(y) = \frac{\Gamma(q\sigma+1)}{\sigma\Gamma(q+1)} y^{q-1-1/\sigma} f_{\sigma}(y^{-1/\sigma}), \tag{5}$$

where f_{σ} is the density function of a positive σ-stable random variable. Moreover, the random variables $B_{j+\theta/\sigma, n/\sigma-j}$ and $S_{\sigma,(\theta+n)/\sigma}$ are assumed to be independent. When interest is in estimating the number of rare species, then one can resort to the cumulated estimator

$$\hat{M}_{n,m}(1, \ldots, \tau) = \sum_{i=1}^{\tau} \hat{M}_{n,m}(i). \tag{6}$$

More generally, if $\{l_1, \ldots, l_\tau\}$ are distinct integers such that $l_i \in \{1, \ldots, m+m\}$ for every i, then the number of species that appear with frequency l in $\{l_1, \ldots, l_\tau\}$ in the enlarged sample of size $n+m$ can be estimated by

$$\hat{M}_{n,m}(l_1, \ldots, l_\tau) = \sum_{i=1}^{\tau} \hat{M}_{n,m}(l_i). \tag{7}$$

Finally, [1] derived closed form expressions for the joint conditional distribution of vectors $(M_{n,m}(l_1), \ldots, M_{n,m}(l_\tau))$, with all the l_i's being distinct indexes in $\{1, \ldots, n+m\}$. This allows us to prove that,

$$\frac{M_{n,m}(l_1, \ldots, l_\tau)}{m^{\sigma}} \xrightarrow{\text{w}} \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)_{l_i-1}}{l_i!} \right) S_{\sigma,\theta,n,j}. \tag{8}$$

2 Methodology

While deriving the closed form expressions in (2) and (7), [4] did not consider the problem of associating a measure of uncertainty to $\hat{M}_{n,m}(l)$ and $\hat{M}_{n,m}(l_1,\ldots,l_\tau)$. We present a novel methodology to approximately quantify the uncertainty of the estimators in (2) and (7).
To this end, fluctuations (3) and (8) provide a useful tool for approximating the distribution of the random probabilities $M_{n,m}(l)$ and $M_{n,m}(l_1,\ldots,l_\tau)$. Specifically, we aim at exploiting such limiting distribution in order to construct asymptotic credible intervals for the estimators.

First, we observe that the same limiting results would clearly hold true for any scaling factor $r(m)$ such that $r(m)\approx m^\sigma$. Numerical investigations show that, as soon as θ and n are not overwhelmingly smaller than m, the asymptotic estimator

$$\hat{M}'_{n,m}(l) = m^\sigma \frac{\sigma(1-\sigma)_{l-1}}{l!}\mathbb{E}[S_{\sigma,\theta,n,j}] \tag{9}$$

can be far from the exact estimator $\hat{M}_{n,m}(l)$. For this reason we introduce the scaling $r^*(m,l)\approx m^\sigma$ such that $\hat{M}_{n,m}(l) = r^*(m,l)(\sigma(1-\sigma)_{l-1}/l!)\mathbb{E}[S_{\sigma,\theta,n,j}]$, and we define the unbiased estimator

$$\hat{M}^*_{n,m}(l) = r^*(m,l)\frac{\sigma(1-\sigma)_{l-1}}{l!}\mathbb{E}[S_{\sigma,\theta,n,j}]. \tag{10}$$

Similar observations hold true for the estimator $\hat{M}_{n,m}(l_1,\ldots,l_\tau)$. According to (8), the asymptotic counterpart of this estimator coincides with $\hat{M}'_{n,m}(l_1,\ldots,l_\tau) = \sum_{1\le i\le\tau}\hat{M}'_{n,m}(l_i)$. In particular, we introduce the scaling $r^*(m,l_1,\ldots,l_\tau)\approx m^\sigma$ such that

$$\hat{M}_{n,m}(l_1,\ldots,l_\tau) = r^*(m,l_1,\ldots,l_\tau)\sum_{i=1}^{\tau}\left(\frac{\sigma(1-\sigma)_{l_i-1}}{l_i!}\right)\mathbb{E}[S_{\sigma,\theta,n,j}]$$

and we define the unbiased estimator

$$\hat{M}^*_{n,m}(l_1,\ldots,l_\tau) = r^*(m,l_1,\ldots,l_\tau)\left(\sum_{i=1}^{\tau}\frac{\sigma(1-\sigma)_{l_i-1}}{l_i!}\right)\mathbb{E}[S_{\sigma,\theta,n,j}]. \tag{11}$$

To keep the exposition as simple as possible we do not provide the expression for the factors $r^*(m,l)$ and $r^*(m,l_1,\ldots,l_\tau)$. See [3] for a similar approach in the context of Bayesian nonparametric inference for the number of new distinct species generated by the additional sample.

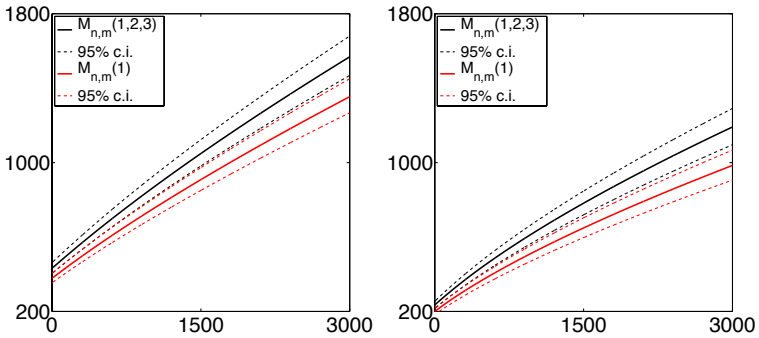
The strategy we follow, in order to obtain asymptotic credible intervals for $\hat{M}_{n,m}(l)$ and $\hat{M}_{n,m}(l_1,\ldots,l_\tau)$, starts with evaluating appropriate quantiles of the distribution of the limiting random variable $S_{\sigma,\theta,n,j}$. For instance let s_1 and s_2 be quantiles of the distribution of $S_{\sigma,\theta,n,j}$ such that (s_1,s_2) is the 95% credible interval with respect to this distribution. Then, according to (3) and (10), the set

$$\left(r^*(m,l)\frac{\sigma(1-\sigma)_{l-1}}{l!}s_1, r^*(m,l)\frac{\sigma(1-\sigma)_{l-1}}{l!}s_2\right) \tag{12}$$

is a 95% asymptotic credible interval for $\hat{M}_{n,m}(l)$. Analogous observations hold true for the estimator $\hat{M}_{n,m}(l_1,\ldots,l_\tau)$. In order to determine the quantiles s_1 and s_2, we devised an algorithm for sampling the limiting random variable $S_{\sigma,\theta,n,j}$ that involves sampling from the distribution (5). To this end, we combine the algorithm proposed in [3] with the so-called fast rejection algorithm for sampling from an exponentially tilted positive σ-stable random variable (see [5]).

3 Illustration

In order to illustrate the proposed methodology we consider two cDNA libraries of the amitochondriate protist *Mastigamoeba balamuthi*. The two libraries differ since one is non-normalized, whereas the other one is normalized, namely it undergoes a normalization protocol which aims at making the frequencies of genes in the library more uniform so to increase the discovery rate. See [9]. Due to the high cost of such protocols, an accurate estimate of the number of rare species in an unobserved sample can be of great importance in deciding whether it is worth applying them. For the *non-normalized Mastigamoeba* dataset the observed sample consists of $n = 715$ ESTs with $j = 460$ distinct genes whose frequencies are $m_{i,715} = 378,33,21,9,6,1,3,1,1,1,1,5$ with $i \in \{1,2,\ldots,10\} \cup \{13,15\}$. For the *normalized Mastigamoeba* dataset the observed sample consists of $n = 363$ with $j = 248$ distinct genes whose frequencies are $m_{i,363} = 200,21,14,4,3,3,1,0,1,1$ with $i \in \{1,2,\ldots,9\} \cup \{14\}$.



(a) Non-normalized Mastigamoeba (b) Normalized Mastigamoeba

Fig. 1: Mastigamoeba libraries. Exact estimates $\hat{M}_{n,m}(1)$ (red solid curves) and $\hat{M}_{n,m}(1,2,3)$ (black solid curves) together with asymptotic 95% credible intervals (dashed curves). The size m of the additional sample ranges in $[0,3000]$.

As for our analysis, we call a species unique if it has frequency 1 and consider rare those species that have abundance not greater than $\tau = 3$. For both datasets, after setting the parameters θ and σ by means of the empirical Bayes procedure suggested in [3], we compare the estimated numbers of rare and unique species, that is $\hat{M}_{n,m}(1)$ and $\hat{M}_{n,m}(1,2,3)$, for a size m of the additional unobserved sample in $[0,3000]$. By applying the introduced methodology we endow the exact estimate with asymptotic credible intervals. Figure 1 suggests that, for the non-normalized library, the estimated number of both rare and unique species grows slightly faster than the same quantity for the normalized library. Moreover, by inspecting Figure 1(a) and Figure 1(b), it is apparent that most of the rare species are unique. This fact is confirmed by the estimates in Table 1 and holds true in both the observed and, according to the estimators $\hat{M}_{n,m}(1)$ and $\hat{M}_{n,m}(1,2,3)$, the enlarged sample.

Table 1: Mastigamoeba libraries. Number of observed unique (m_1) and rare ($m_1+m_2+m_3$) species; estimated number of unique ($\hat{M}_{n,3000}(1)$) and rare ($\hat{M}_{n,3000}(1,2,3)$) species after an unobserved sample of size 3000; corresponding 95% credible intervals.

	m_1	$\hat{M}_{n,3000}(1)$	95% c.i.	$m_1+m_2+m_3$	$\hat{M}_{n,3000}(1,2,3)$	95% c.i.
Mast.	378	1354.4	(1268.0,1450.7)	432	1568.7	(1468.7,1680.2)
Mast. norm.	200	984.8	(906.0,1067.2)	235	1191.4	(1096.0,1291,1)

Acknowledgements The authors are also affiliated to the Collegio Carlo Alberto in Moncalieri, whose support is acknowledged. The first author is also supported by the European Research Council (ERC) through StG "N-BNP" 306406.

References

1. Cesari, O., Favaro, S. and Nipoti, B.: Posterior analysis of rare variants in Gibbs-type species sampling models. Preprint (2012).
2. Elandt-Johnson, R. C.: Probability models and statistical methods in genetics, John-Wiley and Sons Inc. (1971).
3. Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I.: Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. J. Roy. Statist. Soc. Ser. B **71**, 993–1008 (2009).
4. Favaro, S., Lijoi, A. and Pruenster, I.: Conditional formulae for Gibbs-type exchangeable random partitions. The Annals of Applied Probability **23**, 1721–1754 (2013).
5. Hofert, M.: Efficiently sampling nested Archimedean copulas. Comput. Statist. Data Anal. **55**, 57–70 (2011).
6. Lijoi, A., Mena, R.H. and Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika **94**, 769-786 (2007).
7. Perman, M., Pitman, J. and Yor, M.: Size-biased sampling of Poisson point processes and excursions. Probab. Theory Related Fields **92**, 21–39 (1992).
8. Pitman, J. and Yor, M.: The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997).
9. Susko, E. and Roger, A.J.: Estimating and comparing the rates of gene discovery and expressed sequence. Bioinformatics **20**, 2279–2287 (2004).